阿里云天池数据集——

2012-2015 淘宝天猫母婴用品销售数据分析

1. 项目数据集介绍

本样例数据来自阿里云-天池数据集的淘宝和天猫上婴儿用品销售信息（，本数据集包括 2 个文件，分别为：

- 表 1：(sample)sam_tianchi_mum_baby_trade_history.csv——婴儿用品交易记录，导入数据后命名表为 trade；
  其中包含 29972 行，7 列。其中 7 个字段命名和含义如下：

user_id	用户 ID（BIGINT）
auction_id	购买行为编号(BIGINT)
cat1	商品所属大类编号（BIGINT）
cat_id	商品 id，以下简称商品所属小类编号，是 cat1 的子类（BIGINT）
property	商品属性，此处无用已删除
buy_mount	商品销量（个数）（INT）
day	购买日期（DATE）

- 表 2：(sample)sam_tianchi_mum_baby.csv——部分婴儿信息(900 余条)，导入数据后命名为 baby。
  其中包含：954 行，3 列。3 个字段含义如下：

user_id	用户 ID（BIGINT）
birthday	用户出生日期（DATE）
gender	用户性别 （0 男孩 1 女孩 2 性别不明）（INT）

2. 提出问题&分析思路

根据 2012 年 7 月到 2015 年 2 **月中旬的销售数据和部分，由于此处不知道商品单价，暂时不考虑销售额，本次分析核心目的为提高销量**。本数据中可以利用的分析指标有：

营运指标中成交指标→成交数量和成交用户数；会员指标中会员复购率和平均购买次数；以及用户的相关信息。

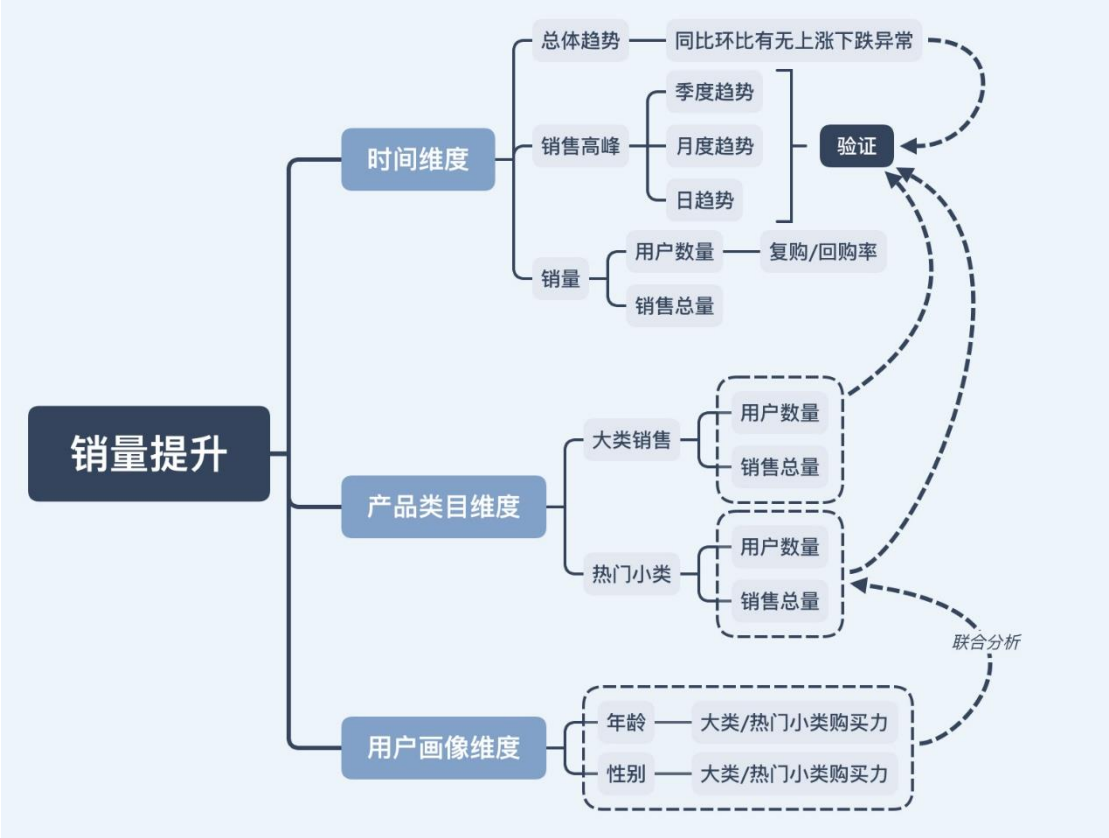
由此提出一下问题：

1） 销量、订单数在时间上的整体趋势如何，按月度、季度等销售情况如何，是否有规律性峰值，是否有下降趋势和销售异常低谷，如果有则分析原因；
2） 用户复购率如何，客户组成的新老客比例如何，销量贡献是由新客还是老客提供；
3） 各大类商品的具体销售情况如何，销售比较突出的商品小类是什么，有哪些商品可以重点进行营销；

4） 对用户进行画像，探究不同性别不同年龄的用户购买情况如何；

5） 对热门商品进行购买趋势分析和用户画像分析，提出更有针对性地营销建议。

下面根据时间、商品品类、用户画像三个维度进行分析。分析思路如下。



3. 数据初步分析&清洗

对婴儿信息进行分析，发现有 1984 年出生的明显不合理值，去掉不合理值，得到婴儿年龄约为 0-10 岁（购买数据截止到 2015 年初）。

	birthday
平均	20118187
标准误差	676.523
中位数	20120828
众数	20130309
标准差	20873.78
方差	4.36E+08
峰度	1.785105
偏度	-1.17199
区域	129610
最小值	20021205
最大值	20150815
求和	1.92E+10
观测数	952

对每单购买量进行分析，结果显示有异常值出现，数据偏度较大。对此进行数据清洗。

buy_mount	
平均	2.544126
标准误差	0.369607
中位数	1
众数	1
标准差	63.98688
方差	4094.321
峰度	20133.78
偏度	133.4585
区域	9999
最小值	1
最大值	10000
求和	76250
观测数	29971

考虑到主要分析主体对象为家庭购买单位、自用非盈利性质购买，去掉大额订购订单。考虑到婴儿用品日期要求严格且婴儿成长、需求变化快，不能大量囤货，以半年为极端囤货周期。

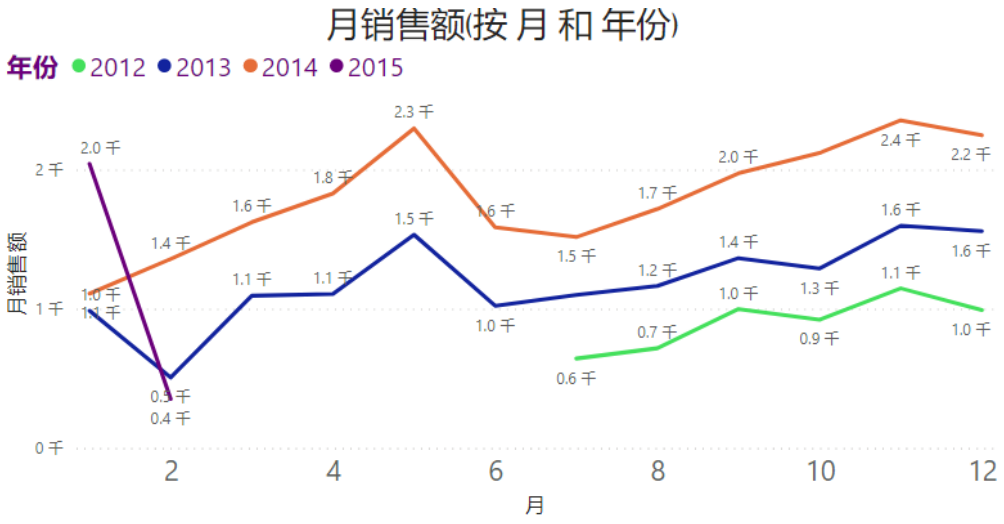
以奶粉为例进行估算。

考虑到 1 段奶粉消耗量最大，以婴儿一段奶粉用量为估算标准。以婴儿每公斤体重每月需要 500g 奶粉、婴儿 3-6 个月体重从 3kg 到 10kg 线性增长进行估算，婴儿一段奶粉大概需要 50 罐/半年（以市场上较小分量包装 400g/罐为准）。结论为去掉购买数量大于 50 罐的数据。共去掉 91 个异常值。经检查无缺失值和其他异常值。

```
DELETE FROM data.trade
WHERE buy_mount > 50;
```

4. 时间维度分析

1） 首先确认销量走势：

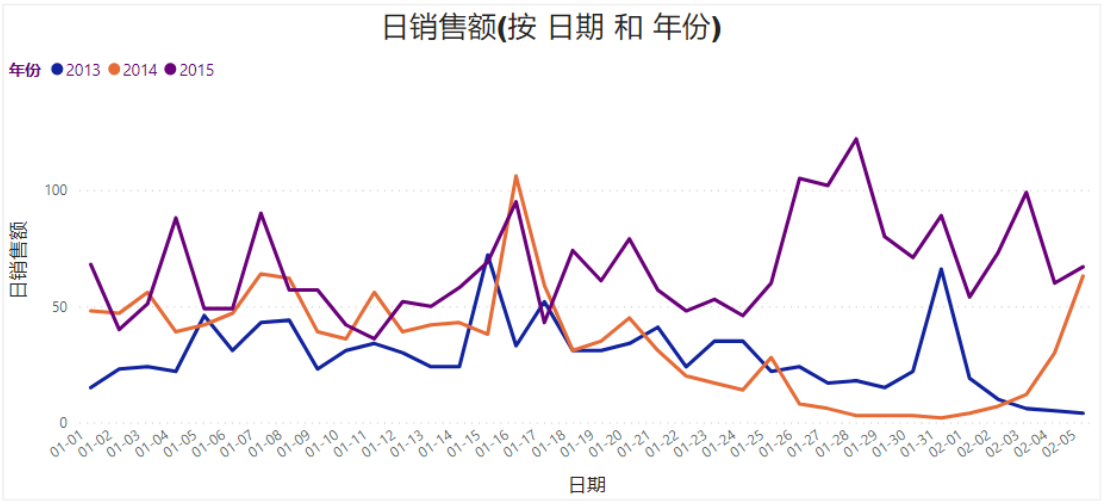


按月可视化，发现 2015 年的月均销量陡降，其余年份按年环比增长且趋势稳定。其中 5 月和 11 月、12 月为销售的高峰期。

下面来探究 2015 年的销量下降原因。

```
SELECT
    year(day),day(day),date_format(day,'%m-%d'), sum(buy_mount)
    FROM data.trade
WHERE
     (month(day) in (1,2)) and (date_format(day,'%m-%d') < '02-06')
GROUP BY
    day
ORDER BY
    date_format(day,'%y-%m-%d');
```

按日绘制销量曲线，与 2013 年、2014 年同期相比如图：

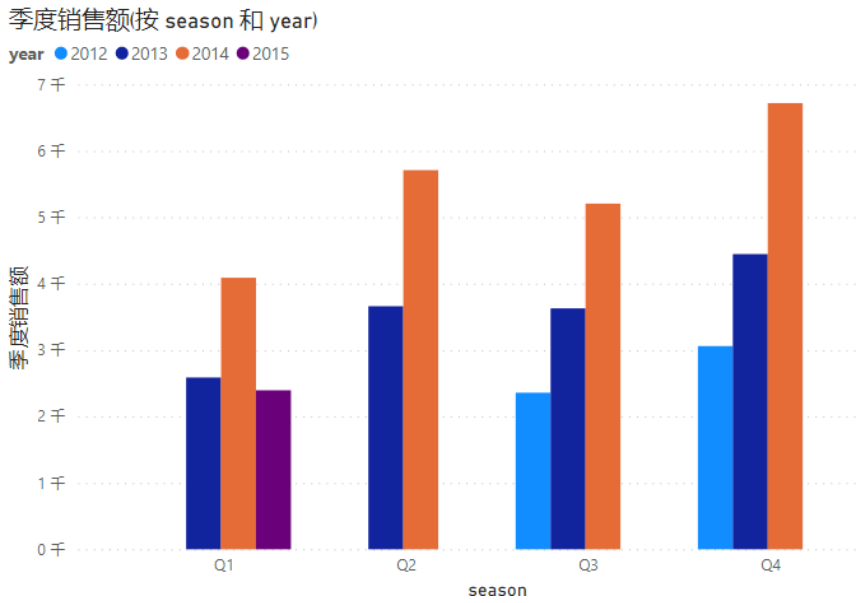


可以看出 15 年的销量与 13、14 年相比并无明显差异，峰值也都有相似波动，而且 15 年在 1 月末 2 月初的销量比 13、14 年同比显著增长。可以得出结论：**15 年的销量明显下降是由数据缺失导致。**

2） 其次进行季度销量的比较：

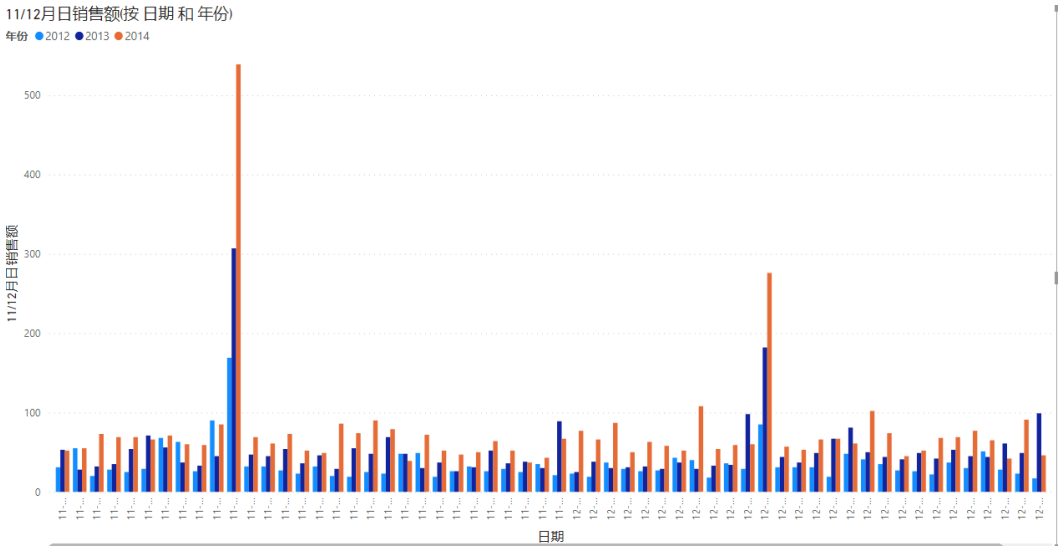
```
SELECT
    case
    when month(day) in (1,2,3) then 'Q1'
    when month(day) in (4,5,6) then 'Q2'
    when month(day) in (7,8,9) then 'Q3'
    when month(day) in (10,11,12) then 'Q4'
    end as season,
    year(day) as year, day, sum(buy_mount)
    FROM data.trade
GROUP BY
    year(day), season
ORDER BY year, season
    ;
```

season	year	day	sum(buy_mount)
Q3	2012	2012-08-16	2358
Q4	2012	2012-12-23	3060
Q1	2013	2013-01-02	2587
Q2	2013	2013-04-18	3660
Q3	2013	2013-07-25	3628
Q4	2013	2013-10-31	4445
Q1	2014	2014-03-29	4089
Q2	2014	2014-06-19	5709
Q3	2014	2014-08-02	5205
Q4	2014	2014-11-01	6719
Q1	2015	2015-02-02	2394

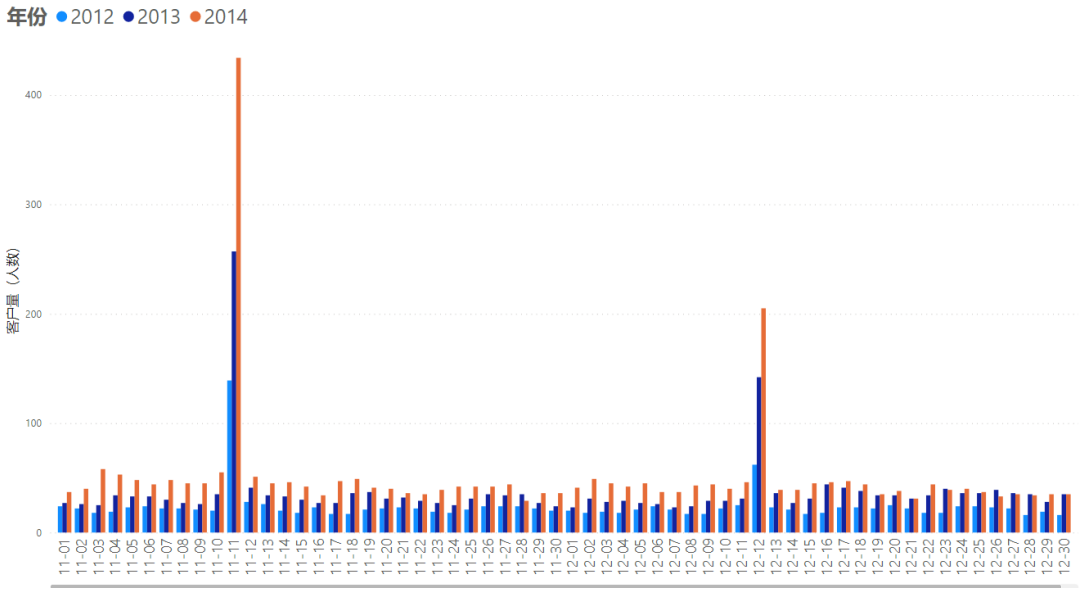


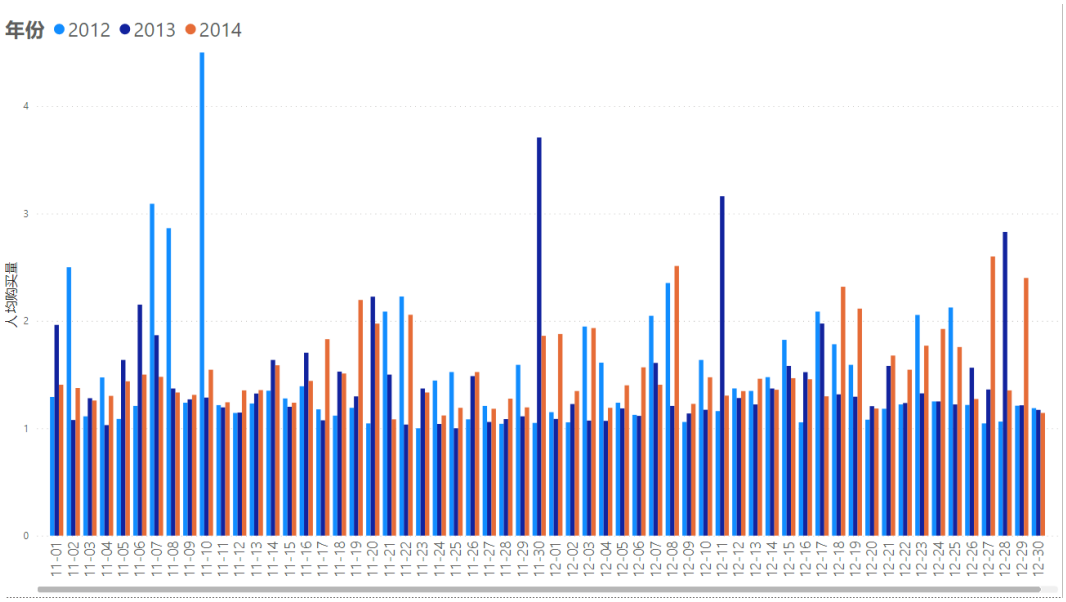
从图中可以看出，季度销量逐年同比上升，其中季度销量表现呈现第四季度最优趋势。怀疑是双十一、双十二等购物节影响或季度性发售的产品影响。

在时间维度中，首先按日画出第四季度销量的曲线。



可以看出，在双十一、双十二当天都产生了一个销量峰值。此时产生疑问：是客户增多还是客户的购买量增多了呢？对此进行针对人均购买量和用户量的时间维度可视化。





可以看出，人均购买量无明显变更，然而客户人数却在‘双十一’、‘双十二’当天有很大提升，并且逐年增长。可以得出购物节主要是促进更多用户进行消费。值得注意的是在双十一双十二购物节前一天或几天往往有人均购买量微增的趋势，经查在此期间有个别用户大量购买的行为，怀疑是同行或某些极个别囤货行为，此处缺失数据，可以进一步调查并做好相应备货准备。

为增加购物街高峰客户人数，分析高峰期客户新老客比例，发现历年购物节期间仅有一到两名老客下单，由此计算复购率，发现每年复购率极低。需要更多数据进行客户流失原因的数据分析。

user_id	auction_id	cat_id	cat1	buy_mount	day
57700536	19798005819	50010549	50008168	1	2013-12-12

user_id	auction_id	cat_id	cat1	buy_mount	day
116466705	16949927732	50010549	50008168	1	2013-11-11
21833701	35304106898	50012359	122650008	1	2013-11-11

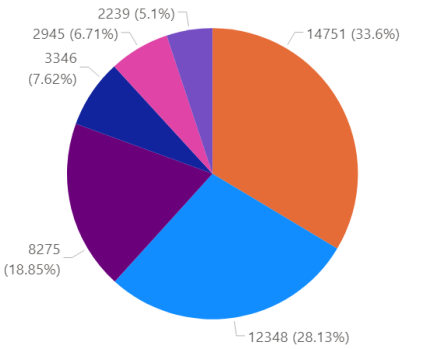
观察到，即使时间尺度拉到年为单位，复购率依然极低。

下单年	下单人数	当年复购人数	当年复购率
2012	3585	0	0.00%
2013	9713	6	0.06%
2014	14339	4	0.03%
2015	1435	2	0.14%

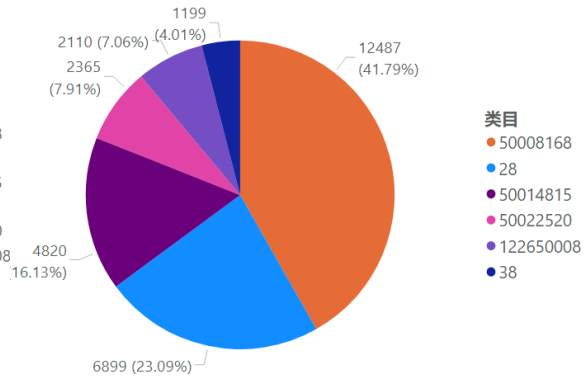
5. 商品品类维度分析

共有六个商品大类：

各大类总销量

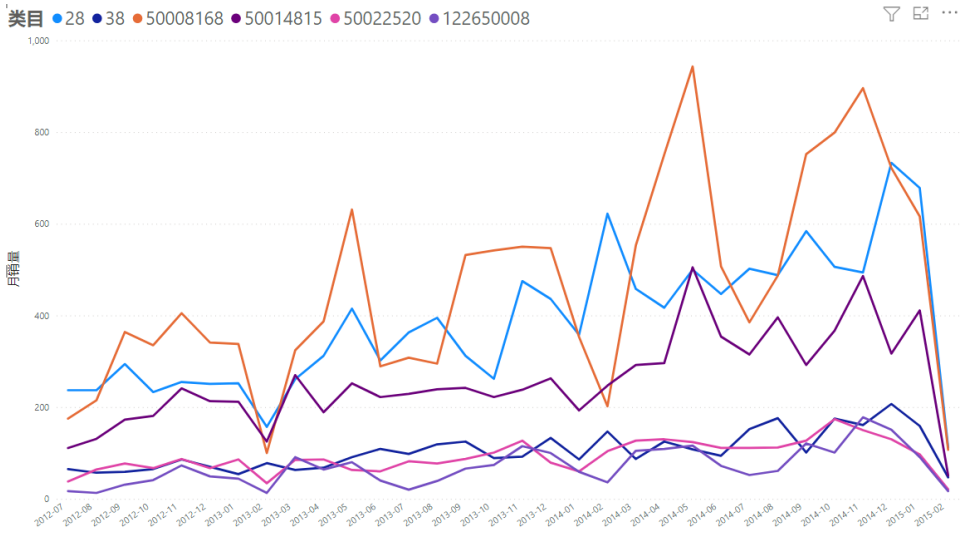


各大类购买总人次

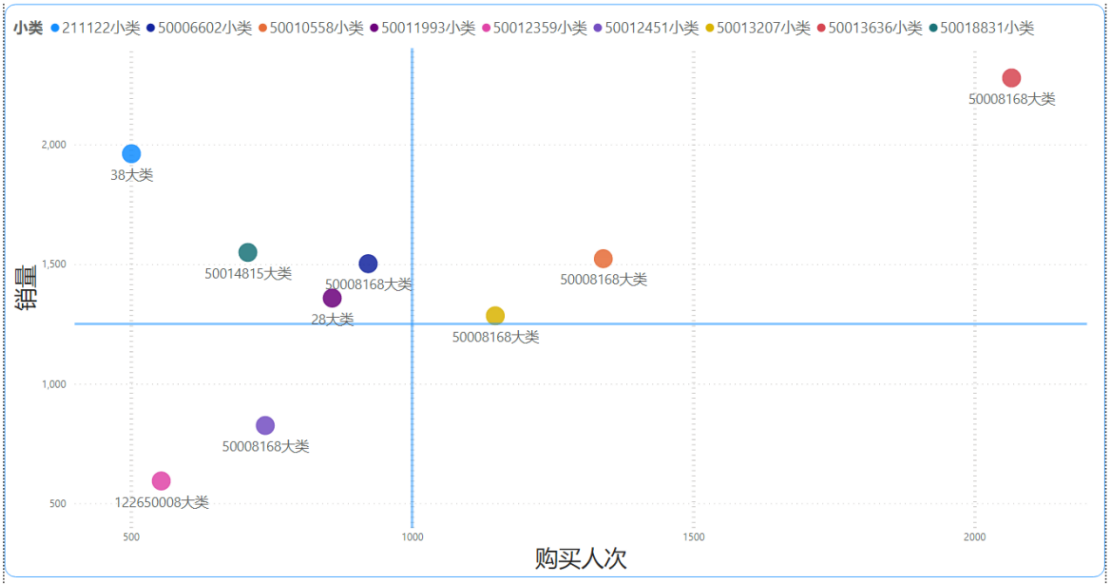


可以看出类目 50008168 在两项比较中都是居于首位，类目 28 和 50014815 居于其次。此时不知道具体商品为何物以及其单价，所以不从销售额的角度衡量。仅从目前结果来看，推广和促销活动、库存储备应集中在三个较高的大类上。

其次看热门大类销量随时间变化曲线，可以看出每年进入春夏季和双十一双十二期间都会有销量的大幅提升，且提升主要集中在三个热门大类。值得注意的是，最热门的大类 50008168 会在五月、十一月出现销量的骤增，怀疑季节性产品、季节性促销活动或其他，需要进行其他维度的数据分析来探究原因，此处提示注意库存储备。

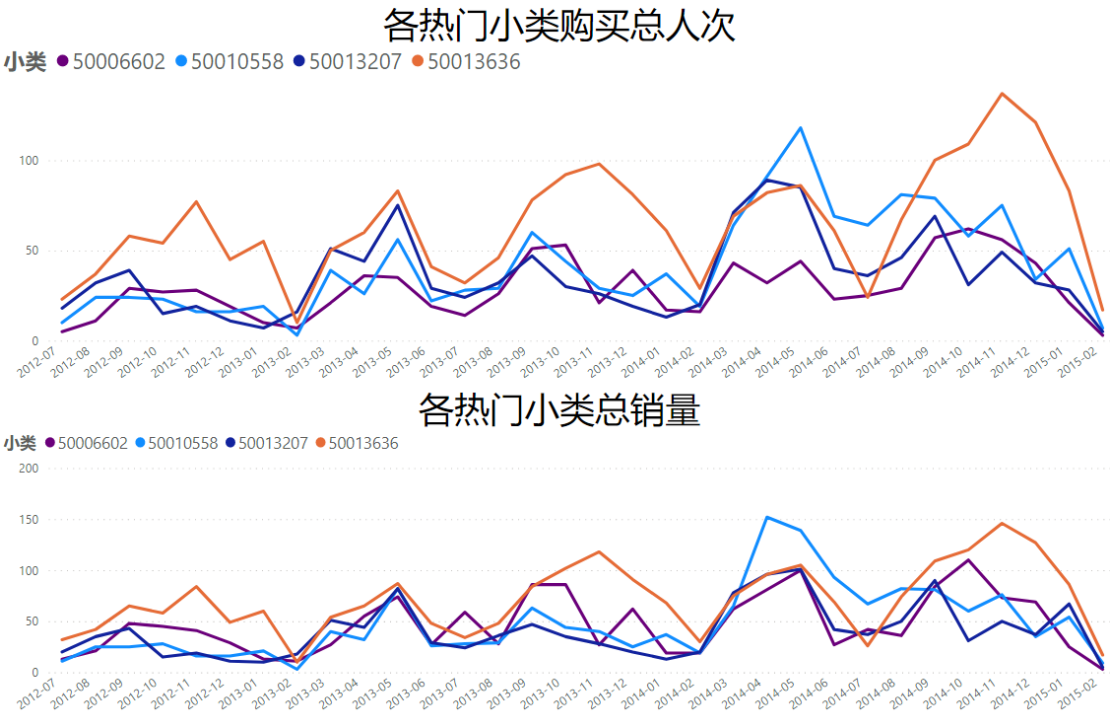


接下来分析小类商品的销售情况。由于小类商品数量较多，采取二八法则，定义热门商品（销量和购买人次均为小类前十名），结果按照销量和购买人次采用象限法进行分析。如图所示。

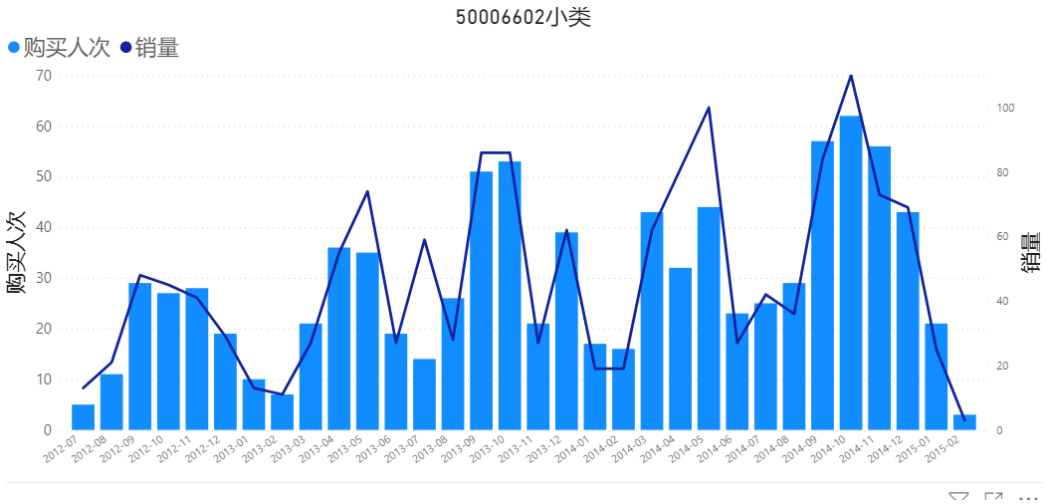


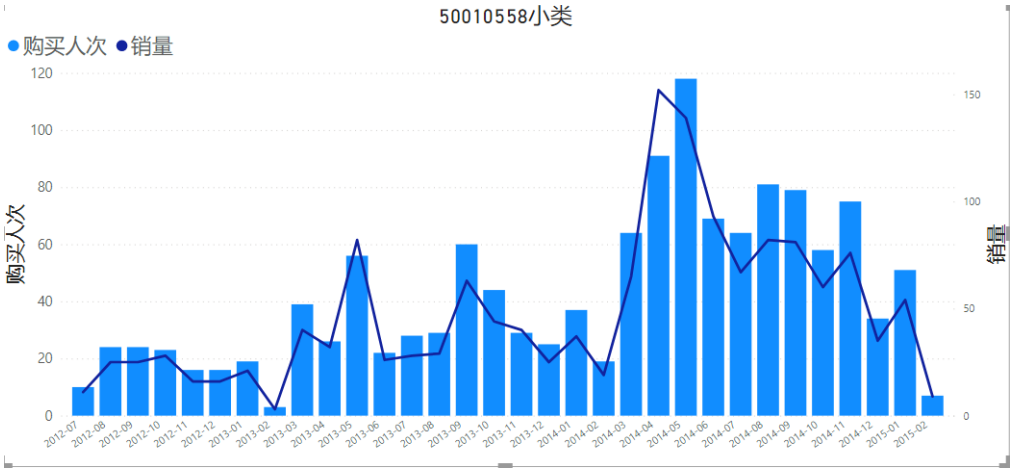
可以看出最优的四个小类皆属于 50008168 大类，基于此，对 50008168 大类中的四个产品进行举例式分析。

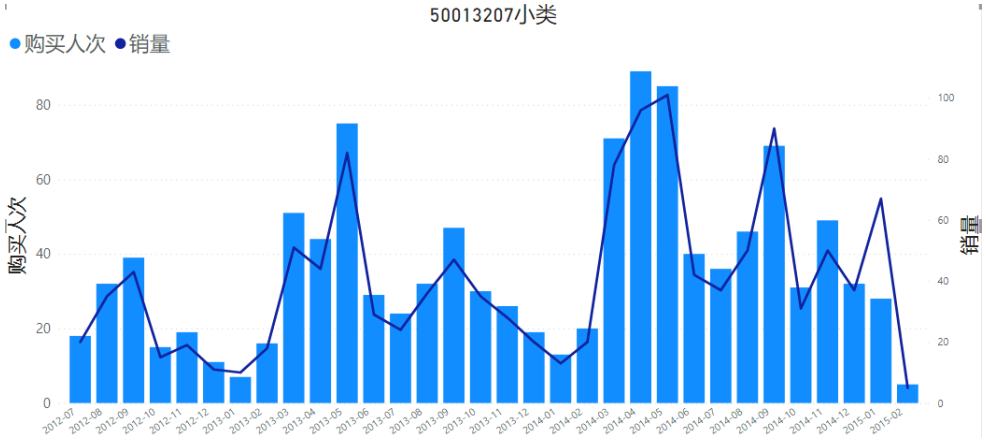
首先看销售趋势，发现冠军产品 50013636 是最主要出现五月、十一月峰值的，50010558 次之。

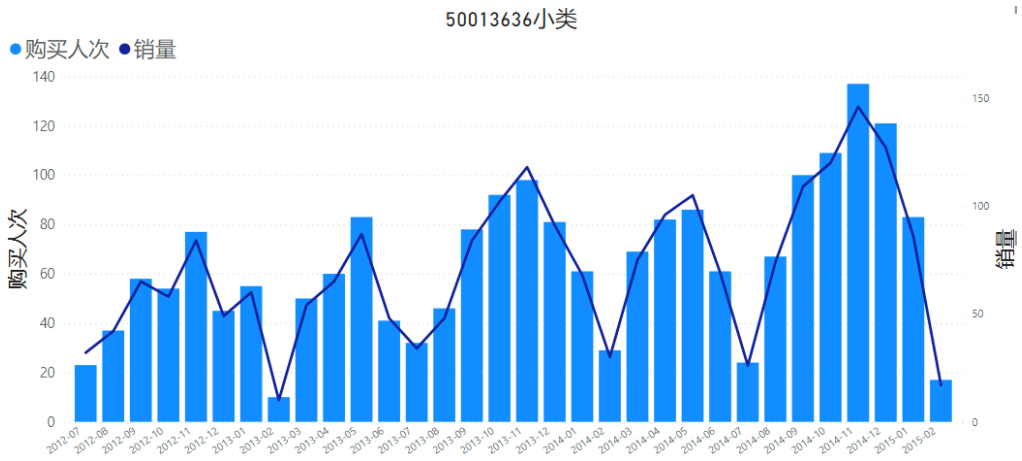


紧接着是小类的销量、购买人次分析，发现仅有 50006602 小类容易在购物高峰期间出现少量囤货现象。可能是由于商品性质导致。从而也提醒我们对于其他热门商品进行新客户的积极挖掘。


50006602小类
购买人次
销量
购买人次
销量

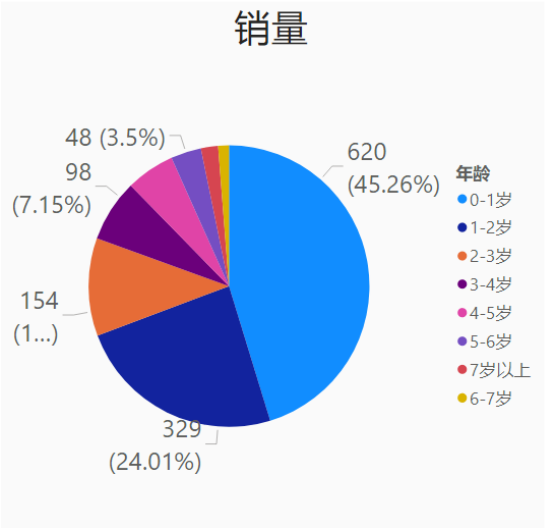

50010558小类
购买人次
销量
购买人次
销量

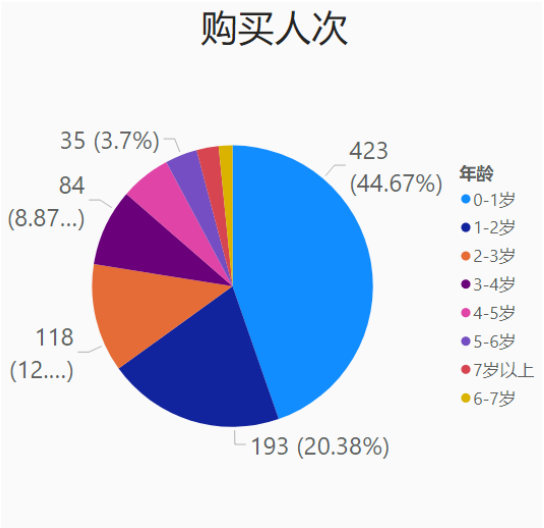

50013207小类
购买人次
销量
购买人次
销量

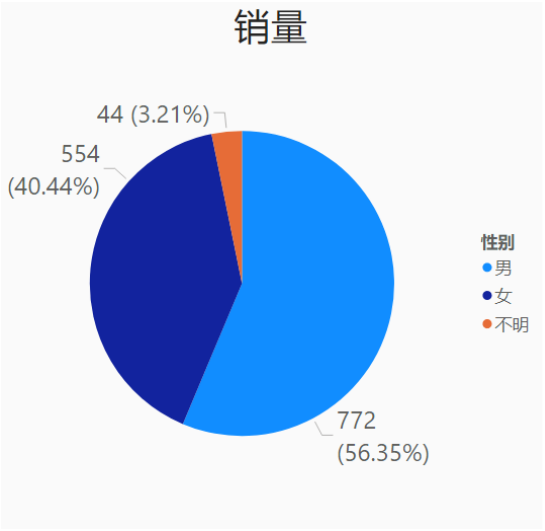


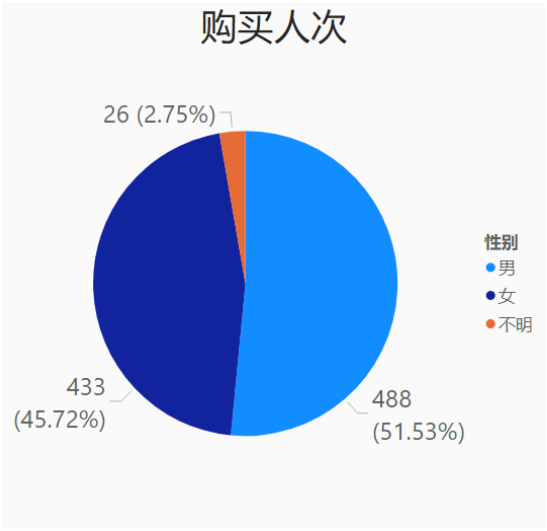
6. 用户画像分析

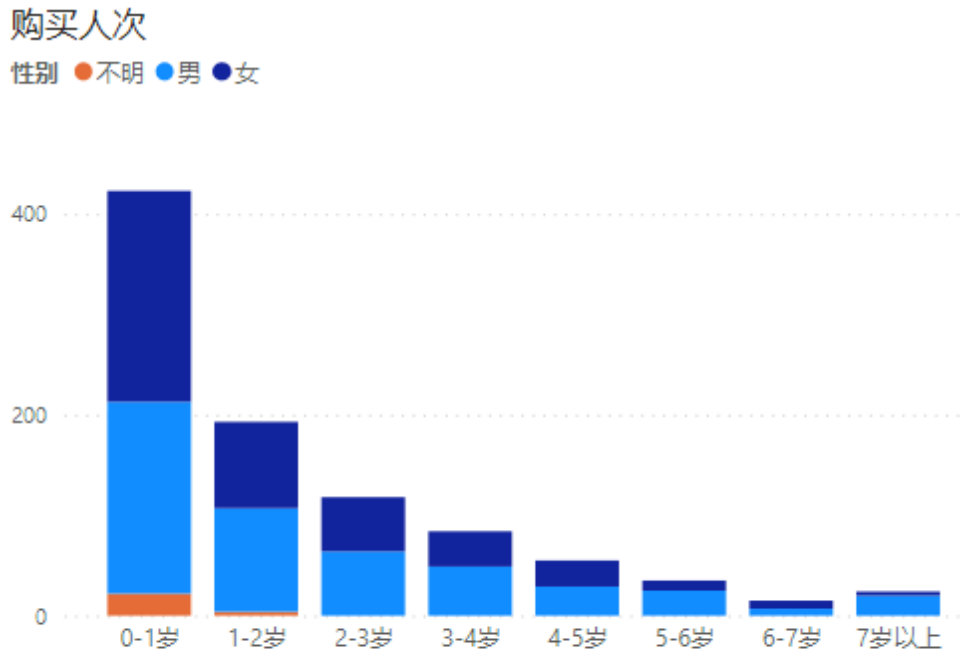
用户按照年龄分析购买量和购买人次，可以看出主要购买生力军还是 0-1 岁的婴幼儿，购买力随年龄增长而递减。而从性别来看，男女宝宝购买人次上并无明显差异的同时，购买量上男宝宝大于女宝宝，说明男宝宝的单次购买量相对大一些，在营销信息推送时可以酌情为男宝宝家长推送量相对较大的优惠项目。

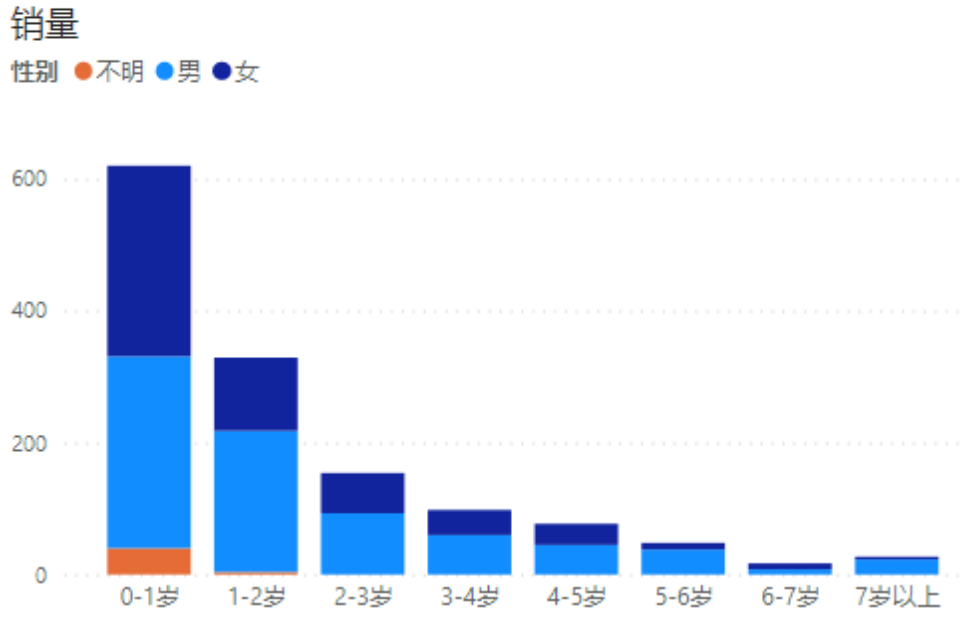




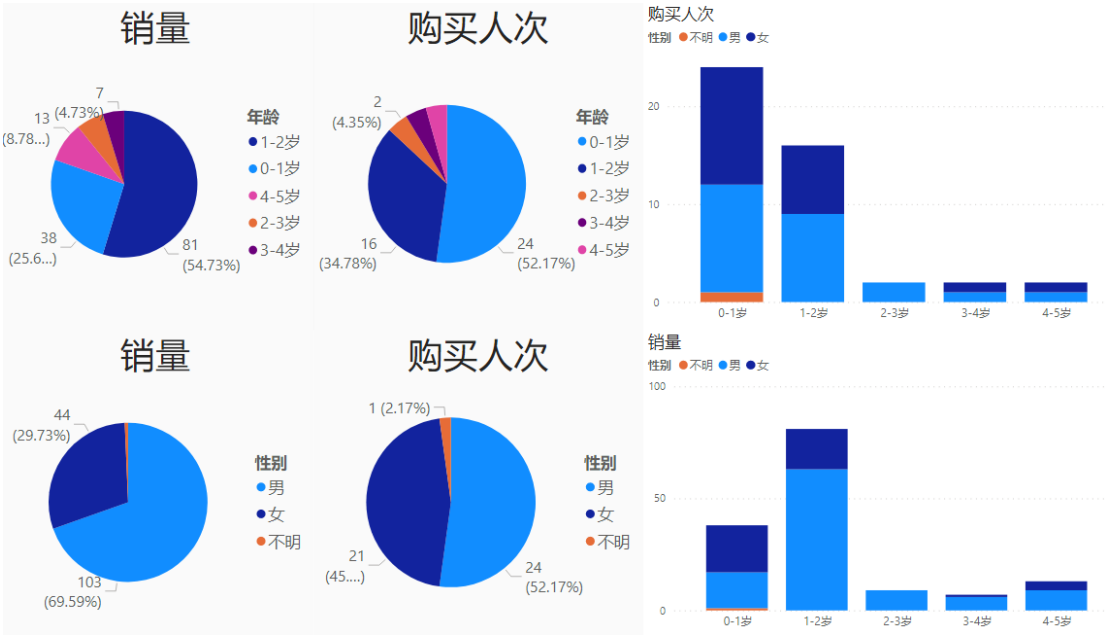

销量
44 (3.21%)
554
(40.44%)
772
(56.35%)
性别
男
女
不明


购买人次
26 (2.75%)
433
(45.72%)
488
(51.53%)
性别
男
女
不明


购买人次
性别 不明 男 女
400
200
0
0-1岁
1-2岁
2-3岁
3-4岁
4-5岁
5-6岁
6-7岁
7岁以上


销量
性别 不明 男 女
600
400
200
0
0-1岁
1-2岁
2-3岁
3-4岁
4-5岁
5-6岁
6-7岁
7岁以上

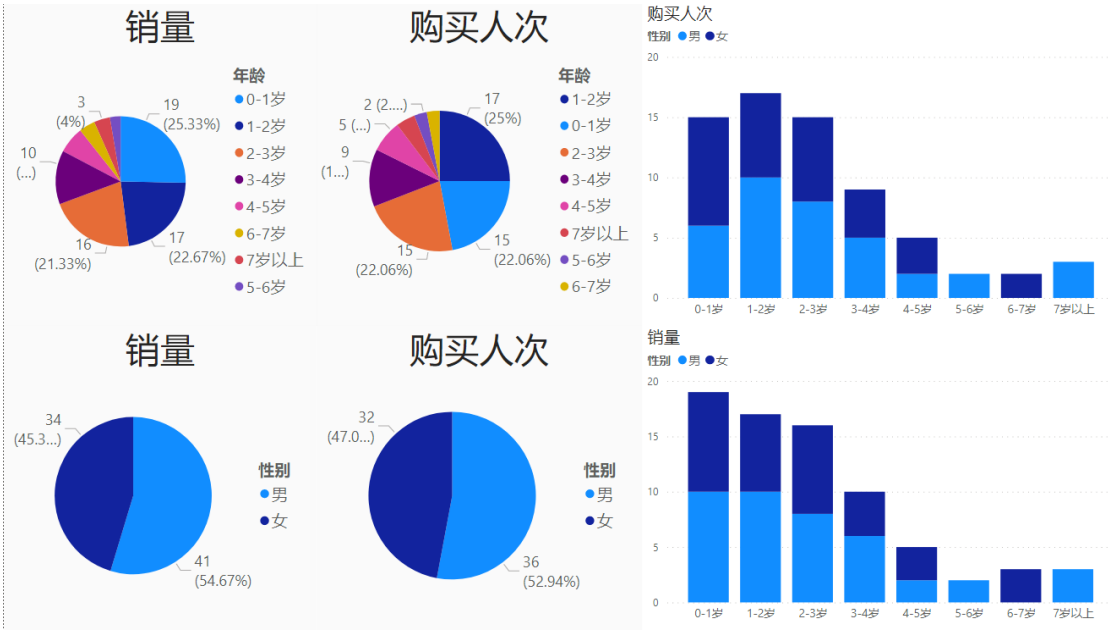
而从柱状图上分析，随着年龄的增长，从 3-4 岁阶段起，女宝宝的购买力明显占优。这说明在留存老客的时候，可以着重对女宝宝家长的宣推工作，也说明对于年龄较大的宝宝家长可以优先对女宝宝进行相关营销活动。



对大类的分析中，得出的年龄、性别可能都是产品本身显而易见的特性，此处不再赘述。特别的，当对 38 大类进行分析时可以得出此处是男宝宝单人购买量高于女宝宝的原因，且此处 1-2 岁宝宝的单人购买量远高于其他年龄段宝宝，这是该商品的营销契机，应面向 1-2 岁男宝宝着重推广 38 大类产品。但由于 38 大类在本次分析中的定义不是最优销量商品，不做进一步深入分析。

下面对上一节的四个明星产品进行用户画像分析

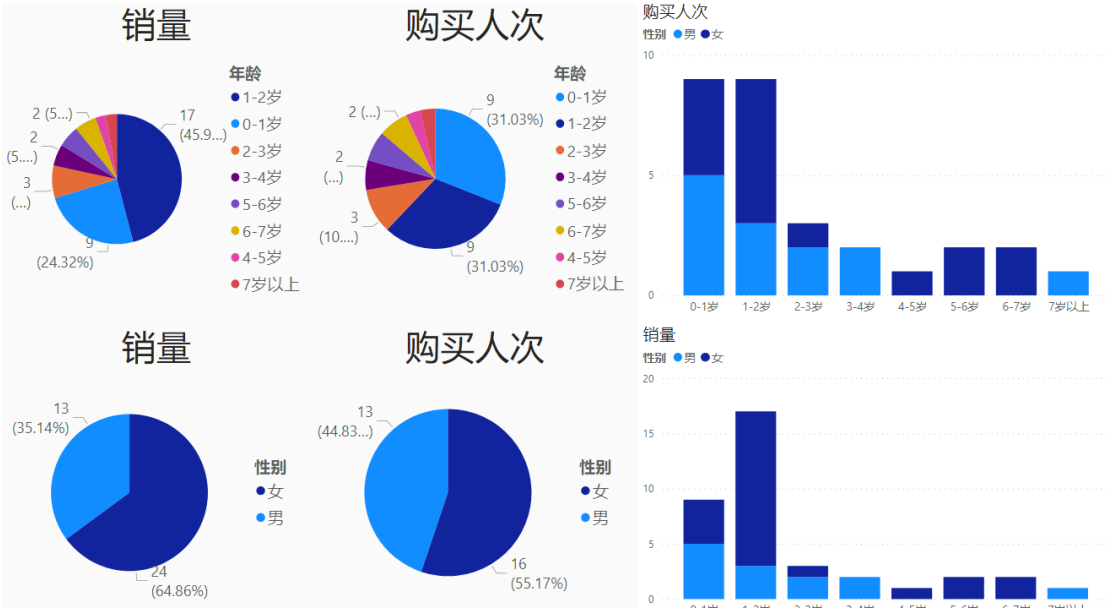
1）50013636 小类



该明星产品主要面向 0-4 岁任意性别的宝宝进行定向宣传。可能是婴幼儿宝宝的某种

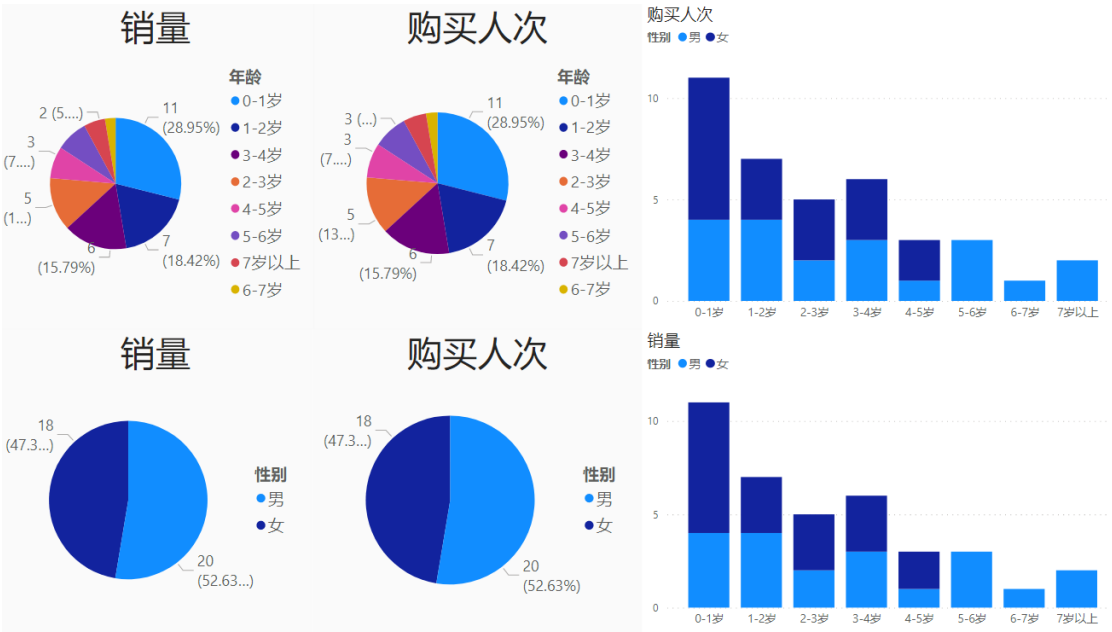
刚需。需要注意此种商品使用时间较长，要加强老客留存，才能有效稳定提升销售额。

2）50010558 小类



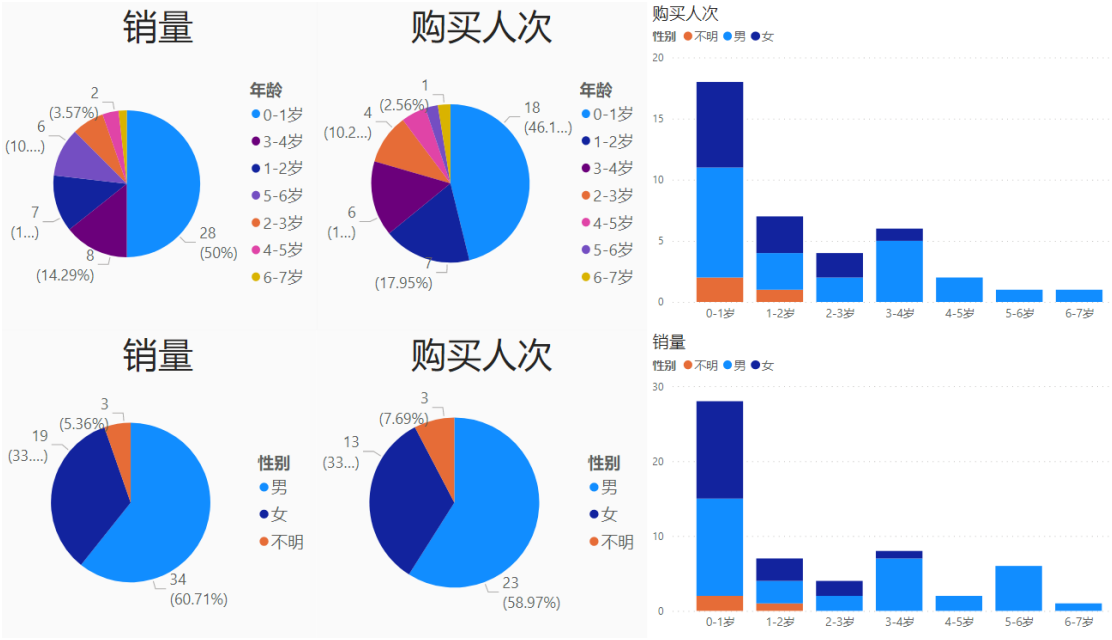
此种商品多为 0-2 岁宝宝使用，尤其应面向女宝宝进行宣传。商品的使用年限不长，如无资源的情况下可以考虑不分配用户回访等老客留存手段。

3）50013207 小类



此商品无明显性别差异，用户年龄段相对较长，可以长期进行推送和用户回访调查、老客优惠等等。

4）50006602 小类



此商品主要面向男宝宝，且年龄主要为 0-1 岁的婴儿，用户年龄段窄，无需过度关注用户留存。值得注意的是该商品销售量较高，若商品单价够高则更需优先考虑此商品的营销工作，尤其是促销节点时的新客获客活动。

7. 结论与建议

结论：

1) 2015 年并非销量陡降而是数据缺失导致曲线下降。相反的，2015 年初销售情况较往年比更好，数据相对完整的 2015 年一月份销量较 2014 年同比增长 85%。
2) 整体复购率极低，购物节中客源主要为新客。
3) 单就销量来看，类目 50008168，类目 28 和 50014815 三个产品大类是销量的前三名，占总销量的 80%。

建议：

1) 建议避开春节销售低谷，在 5 月和 11 月开始前逐步进行宣传推广。可分析相关历年宣推活动 ROI 等。
2) 双十一双十二购物节为主要销量增加点，且在此之前会有少量大额订单，需要相应适当备货和逐步宣传。
3) 针对回头客几乎为零的问题，第一要分析复购率低、客户流失率低高原因（由于婴幼儿用品用户成长和需求变化迅速的特殊性，不排除特殊原因，比如产品本身、价格、竞品、市场环境、行业规律等。因为不知道类别编号对应的实际商品，也不知道关于价格或其他更详细的交易信息，我们无法具体分析）；第二可以暂时先把总体宣传重点放在吸引新客户上，以获得最高的投资回报率。
4) 对于占据销量前 80%的热门大类：类目 50008168，28 和 50014815 给与特别关注，有针对性地进行推广和营销。尤其注意 50008168 会在五月、十一月出现销量的骤增量，需要进行库存的提前预备或可以用预售的形式来为购物高峰做好准

备。

5） 男宝宝的单次购买量相对女宝宝大一些，尤其是 38 大类上，在营销信息推送时可以酌情为男宝宝家长推送量相对较大的优惠项目；女宝宝随着年龄增长在 4 岁以后的购买力高于男宝宝，在留存老客的时候，可以着重对女宝宝家长的宣推工作，也说明对于年龄较大的宝宝家长可以优先对女宝宝进行相关营销活动。

6） 对选出的十个热门小类进行有针对性地营销策略优化。以四个 50008168 大类的产品进行举例分析，结果如下：

小类	销量趋势分析	用户画像分析
50013636	需积极备货预防 5、11 月销量峰值	0-4 岁任意性别的宝宝；商品使用时间较长，要加强老客留存
50010558	也需备货预防 5、11 月销量峰值	0-2 岁宝宝（尤其女宝宝)；商品的使用年限不长，客户留存非必要
50006602	预防购物高峰前的偶发囤货现象	无明显年龄性别差异；可长期推送；客户留存必要
50003207	-	0-1 岁男宝宝；用户年龄段窄，客户留存非必要；商品销售量较高，关注其促销节点时的新客获客